DOCUMENT RESUME

ED 198 145 TM 810 052

AUTHOR Merz, William R.
TITLE Methods of Assessing Bias and Fairness in Tests.
INSTITUTION Applied Research Consultants, Sacramento, Calif.
REPORT NO ARC-TR-121-79
PUB DATE Mar 80
NOTE 59p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Evaluation Methods; Factor Analysis; *Mathematical Models; Statistical Analysis; *Test Bias; Test Construction; Testing Problems; *Test Items
IDENTIFIERS *Test Use

ABSTRACT

Several methods of assessing test item bias are described, and the concept of fair use of tests is examined. A test item is biased if individuals of equal ability have different probabilities of attaining the item correct. The following seven general procedures used to examine test items for bias are summarized and discussed: (1) analysis of variance (ANOVA); (2) transformed item difficulties (TID); (3) item-total score correlations (I-TC); (4) factor analysis (FACTOR); (5) chi-square (CHI); (6) item characteristic curve approaches (ICC); and (7) response fail analysis (RFA). During the test construction phase it is recommended that either the TID, CHI, or three parameter ICC method be employed. Test fairness is defined as the equitable use of test scores in rendering judgments about individuals. Eight methods which use the prediction-regression approach to establish fair test use are discussed. All of these methods demand an outside criterion of success which must be designated and data about performance on that criterion must be collected. A second approach to establish fair test use, which is included in the Equal Employment Opportunity Guidelines and which does not require an outside criterion, is a procedure called the 80 percent rule. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED198145

Applied Research Consultants

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

METHODS OF ASSESSING BIAS AND FAIRNESS IN TESTS

William R. Merz, Ph.D.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. R. Merz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ARC Technical Report #121-79

March, 1980

All Rights Reserved

1979 Applied Research Consultants, Inc.

SUITE 5: 2398 FAIR OAKS BLVD. • SACRAMENTO, CA 95825 • (916) 487-0185
MAILING ADDRESS: P. O. BOX 145 • RANCHO CORDOVA, CA. 95670

TM 810 052

PREFACE

This manual is intended as a general description of the methods for identifying biased test items and for examining tests for fair use. Examples of analyses for test item bias are included. A general example for adverse impact is shown for fair test use.

Much of the material included in this manual is a summary of work done by others. In all instances where the work of another individual has been adapted, a citation crediting that individual has been included. Special credit must be extended to Lawrence M. Rudner, Ph.D. of the National Institute of Education who has collaborated with Dr. Merz in the investigation of test bias and has contributed greatly to the presentation in this manual. Computational examples have been provided by Neal E. Grossen, Ph.D. Critique and comments on the presentation of material has been made by R. A. H. Goodfellow, Ph.D. Typing, organization, and layout have been done by Janice Gorniak.

TABLE OF CONTENTS

	PAGE
PREFACE	1
Table of Contents	2
List of Tables	3
List of Figures	4
SECTION	
PURPOSE	5
INTRODUCTION	5
Terminology Problems	10
TEST ITEM BIAS	13
Analysis of Variance (ANOVA)	16
Transformed Item Difficulty (TID)	18
Point Biserial Correlation (PBIS)	20
Chi-Square (CHI)	21
Factor Analysis (FACTOR)	23
Item Characteristic Curve (ICC)	25
Response Foil Analysis (RFA)	29
Summary	30
Examples Data Set Used	30
Transformed Item Difficulty	34
Chi-Square	35
Three Parameter Item Characteristic Curve	37
FAIR USE OF TESTS	39
Regression Approaches	39
Adverse Impact Approach	43
RECOMMENDATIONS	46
Bias in Test Items	46
Fair Use of Tests	47
Conclusion	49
REFERENCES	50

LIST OF TABLES

NUMBER		PAGE(S)
1	Summary of Selected Item Bias Detection Techniques	14-15
2	Hypothetical Distribution of Distractor Choices	29
3	Total Score Distribution for Examples . .	31-32
4	Listing of Item Difficulties for Each Group	33
5	Hypothetical Application of the 80 Percent Rule No Adverse Impact	43
6	Hypothetical Application of the 80 Percent Rule Adverse Impact	44

LIST OF FIGURES

NUMBER		PAGE
1	Hypothetical ICC	26
2	Bias Solely in Discrimination (a_g) . .	27
3	Bias Solely in Difficulty (b_g) . . .	27
4	Bias in Both Difficulty (a_g) and Discrimination (b_g)	27

PURPOSE

The purpose of this monograph is to describe and illustrate several methods of assessing test item bias and to examine the concept of fair use of tests. It is organized into four parts: an introduction, a description of methods for determining test item bias, a description of methods for assessing fair use of tests, and general recommendations. While test bias and fair test use are problems under all conditions of testing, they will present special difficulties in the competency testing program required of school districts because pupils may be denied diplomas on the basis of test scores. If the items which comprise the test systematically underestimate the performance of any group of individuals, that group is placed at a severe disadvantage and may very successfully challenge the testing program because of the effects of measurement on the underestimation of their achievement. If intact tests cannot be demonstrated to be fair estimators of the achievement of identifiable groups, these groups may successfully challenge the use of these tests in the proficiency program. Therefore, it is very important to determine that the instruments which are used as part of a competency testing program, part of a selection program, or part of an assessment program measure the achievement of all individuals accurately.

INTRODUCTION

In common terms the word bias is perjorative. When an individual uses this term he/she usually means that criteria or standards have been

applied in such a way that he/she has been placed at a disadvantage in being assessed, evaluated, or judged. However, the term bias also has a very special mathematical meaning. When any score or statistic is computed, an estimate of how confident one can be that this score or statistic represents the actual situation can be made. When the score or the statistic has been systematically affected by some unintended or uncontrolled situation and that score or statistic inaccurately represents the true situation, the mathematician refers to this as bias. Technically, then, bias simply means that the number which has been generated to represent an individual's score or some other numerical representation of a situation either overestimates or underestimates what the true value is. Bias, means, mathematically, some systematic error.

Often in testing an individual, who is dissatisfied with her/his score on a test, a rating schedule, an interview, or a checklist, asserts that the test is biased against her/him. Frequently, that person is a member of a group which, on the average, obtains scores lower than other groups. What the person implies is that the test, the rating schedule, the checklist, or the interview systematically underestimates her/his achievement on the trait or the construct which is being measured. The individual claims that the assessment device or the evaluation process somehow discriminates against her/him.

Two separate situations exist when the assertion of bias is made. Merz (1978) has labeled the first situation as "fairness" and the second as "test item bias". Test fairness is defined as the equitable use of

a test score in rendering judgment about individuals of different group measurement. Test item bias is defined as a difference in the probability of obtaining an item correct given equal ability and different group membership. One major difference between the two situations is that fairness addresses the equitable use of intact assessment devices, while bias focuses on specific item performance. Another difference is that bias is usually a concern during the construction of a device, while fairness is addressed when intact instruments are used in the process of making a selection or rendering a judgment.

The recognition that groups differ in levels of performance is as old as the concept of testing itself. Stanley and Hopkins (1975) relate that in the Book of Judges of the Old Testament the pronunciation of the word for ear of corn was used as a test to identify enemy soldiers escaping from defeat. Here a systematic difference between groups in the pronunciation of a word was used as a test in itself. Today a dialectic difference may be viewed as a biasing element if an identifiable group of examinees responds incorrectly to a test item.

The more dependent our educational systems become on using test scores for evaluation, placement, and assessment of proficiency, the more important it is for test developers, test users, and program administrators to examine item performance on specific assessment devices for bias and evaluation procedures for fairness.

The Larry P. v. Wilson Riles et al (Opinion, U. S. District Court for the Northern District of California, 1979) amply illustrates the need for school personnel to attend to the fairness of evaluation systems and the lack of bias in individual test items. In

this case the use of intelligence tests for placing minority children in special education classes has been contested. As a result of that decision, no black child may be placed on the basis of test results which have not been demonstrated to measure accurately for all groups concerned. In a like way, the whole issue of competency testing may rise or fall on the fair assessment of all pupils with test instruments composed of unbiased items so that diplomas and promotions may be awarded in an equitable manner

Equal employment opportunity requires that tests and other selection devices do not place minority individuals at a disadvantage and that they fairly assess the skills required by the job in question and the attainment of those skills by the individual being tested. All in all, it appears that examining test items for bias and the use of tests for fairness will become an ever more important issue as our dependence upon assessment grows.

The National Institute of Education (NIE), underscored the importance and timeliness of this topic by sponsoring a three day conference in December 1975, at Annapolis, Maryland. Experts in the field of testing discussed approaches for constructing less biased items and compared the results of methods of debiasing tests. The fair use of test scores was also examined. The United States Office of Education (USOE) sponsored an invitational conference in May 1976, to examine achievement testing of disadvantaged and minority students who were in programs funded for the improvement of academic attainment. Some of the statistical techniques examined in this manual were presented at or discussed during that conference.

Since those two conferences a great deal of development has taken place in assessing test items for bias. Prior to these conferences a whole set of methodologies labeled item characteristic curve approaches (ICC) were relatively unknown. After those conferences a great deal of work was begun in applying ICC approaches to the area of test item bias. The advantage of this particular set of approaches is that the individual's ability on what is being measured is related directly to the probability of getting an item correct. This is done directly and can be plotted on a scatter plot. It was assumed that this new development would bring clarity to the issue of bias in testing; however, the introduction of the ICC method may have added to the confusion. The major difficulty is that the method is very complicated and requires a higher level of mathematical background than previously employed methods. In addition, there are a number of different experts advocating different uses and different models. In short, ICC approaches have not been around long enough. ICC approaches to test construction, including application to examination of test items for bias, were presented in recent issues of the Journal of Educational Measurement (JEM, 1977) and the Review of Educational Research (RER, 1978). The investigators presently examining ICC approaches are awaiting a new book on latent trait theory (ICC theory) written by Frederick Lord and Melvin Novick scheduled for publication in 1980.

The fair use of test scores in selection was examined in the Spring 1976 issue of the JEM. Equal Employment Opportunity (EEO) Guidelines (1978) detail specific criteria for discovering adverse impact in the selection and other personnel procedures, these criteria apply to differences in test

performance, as well as to proportions of people selected.

Merz and Grossen (1979) and Rudner, Getson, and Knight (1979) have presented comprehensive reviews of item bias detection techniques employing Monte Carlo approaches. Ironson (1978, 1979) and others have investigated actual data. Professional associations interested in testing have addressed the issue of item bias; sessions at the annual meetings of the American Educational Research Association (AERA), the National Council on Measurement in Education (NCME), and the American Psychological Association (APA) in 1978 and 1979 were devoted to methods of examining test items for bias. Educational Testing Service (ETS, 1979) included an entire segment on bias in testing in the materials which were developed on constructing tests for California's competency testing programs. Thus, it is clear that interest in the topics of bias and fairness in testing has been growing.

Terminology Problems

The terminology involved in the areas of bias and fairness is quite confusing. The logic is not clear because terms are often used interchangeably making the concepts which they represent difficult to understand.

More often than not when an individual complains that a measurement device or evaluation procedure was biased, he/she really means that the procedure resulted in a decision which the individual wishes to contest. Often the person alleges that the measurement device or instrument did not accurately measure the attributes being considered. The typical accusation is that a process or an assessment device was biased against

the person or group of persons.

In this sense the individual is saying that the process and the measurement were unfair, that some prejudice was shown, or that factors other than what was really being measured influenced her/his performance and the decision which was made. Here the common use of the term bias (which may be defined with the synonym prejudice) can be equated with the term unfair use. There are really two issues involved here. First, is the data collected with the test instrument systematically affected by attributes other than what is being measured? That is, are individual items or pieces of information biased? Second, in the decision process is the data collected on individuals used in such a way that all individuals have a reasonable opportunity to be judged fairly in the process?

In approaching the questions of bias and fair use, two sets of methodologies are employed. One set is used for situations in which an intact test is administered to members of different groups and these groups attain different mean total scores. Under this condition the usual goal of measurement is to provide data for selecting applicants or for identifying competent individuals. The task is using the test to predict later success in a fair and equitable manner. Here, a statistical technique known as regression analysis is employed using the test under examination as a predictor and some external measure of success as the criterion.

Another set of methodologies is used to identify the items which may systematically differentiate among members of groups. To date these methodologies have not used an external criterion of success and have

focused on single items from a pool which constitutes or will constitute an intact test. The focus of these efforts is usually on constructing a measure which assesses content without introducing systematic error attributable to factors other than those which are the intended object of measurement.

In this monograph a very clear distinction will be made between test item bias and test fairness. The issue of test item bias is problematic because attempts at systematic investigation have been relatively recent. The seven best known procedures have been described by Merz and Rudner (1978), Merz and Grossen (1979), Rudner, Getson and Knight (1979). The clearest discussion of the problem of defining and assessing fairness is found in the Spring 1976 issue of the Journal of Educational Measurement. In it Nancy Petersen and Melvin Novick have organized a detailed evaluation of the models used. Responses to their observations have been made by eminent authorities in the field. In order to deal with the fairness of competency tests extensions, and special applications of the methods discussed will have to be made.

The remainder of this manual, then, will be organized into three sections. The first will deal with test item bias. In this section a general review of the methods that have been used will be presented. Three methods will be presented along with examples, and a list of resources for applying these methods will be presented. In the second section, the issue of fair test use will be described. The limits of exisiting methods will be discussed, and a conceptual presentation of methods will be made. Finally, general recommendations for inspecting test for bias and fairness will be made.

TEST ITEM BIAS

A test item is biased if individuals of equal ability have different probabilities of attaining the item correct. Seven general procedures have been used to examine test items for bias; they are

1. Analysis of Variance (ANOVA),
2. Transformed Item Difficulties (TID),
3. Item-Total Score Correlations (I-TC),
4. Factor Analysis (FACTOR),
5. CHi-Square (CHI),
6. Item Characteristic Curve Approaches (ICC), and
7. Response Foil Analysis (RFA).

These methods have been described extensively by Rudner (1977c), Merz (1978), Merz and Rudner (1978), Merz and Grossen (1979), and Rudner, Getson, and Knight (1979).

All of these methods begin with responses to individual items. Four (TID, I-TC, CHI, and ICC) utilize total scores as well as correct or incorrect responses to items. The table below summarizes the comparison of methods taken from analyses done by Merz and Grossen (1979) and Rudner, Getson, and Knight (1979).

Table 1

SUMMARY OF SELECTED
ITEM BIAS DETECTION TECHNIQUES

METHOD	FOCUS OF ANALYSIS	ASPECT EXAMINED	COMPUTATION EASE	COMPUTATIONAL AIDS NECESSARY
ANOVA	Interaction Between Group Membership and Correct Response	Item Difficulty	Difficult	Computer Necessary
TID	Interaction Between Group Membership and Correct Response	Item Difficulty	Simple	Hand Calculations Possible Calculator Desirable Computer Handy
PBIS	Difference in Item-Total Correlations Between Groups	Item-Total Correlations	Moderate to Simple	Hand Calculation Possible Calculator Desirable Computer Handy
CHI	Difference in Proportions Attaining a Correct Response Across Total Score Categories	Item Difficulty	Moderate to Simple	Hand Calculation Possible Calculator Desirable Computer Handy
FACTOR	Major Loadings on Factors Which Also Have Major Group Membership Loadings	Item Intercorrelation Matrix	Difficult	Computer Necessary
ICC	Difference in Probability of Responding Correctly	Item Difficulty, Item Discrimination, Guessing	Difficult	Computer Necessary
RFA	Difference in Proportions Selecting Distractors	Distractor Difficulty	Simple	Hand Calculation Possible Calculator Desirable Computer Handy

EASE OF INTERPRETATION	MEASURE OF BIAS	SAMPLE SIZE NEEDED	ESTIMATE OF FUNCTION	INVESTIGATORS
Moderate to Difficult	Significance of F for Main Effects and Interactions	Depends on the number of items; Conservative, 100 per group	Not Examined	
Simple	Arbitrary Designation of Distance on Scatterplot	30 to 50 People per Group	Good Good Good Good for Difficulty Poor for Discrimination	Rudner (1977c) Groome, ML (1977) Groome, WR (1977) Merz and Grossen (1979) Rudner, Getson, and Knight (1979)
Simple	Arbitrary Designation of Magnitude of Differences Between Correlations	50 to 100 People per Group	Poor Poor	Rudner (1977c) Merz and Grossen (1979)
Simple	Significance of Chi-Square	100 or More per Group	Satisfactory Satisfactory Satisfactory	Rudner (1977c) Merz and Grossen (1979) Rudner, Getson, and Knight (1979
Difficult	Arbitrary Designation of Magnitude of Factor Loading	Conservative, Number of People 10 Times the Number of Items	Poor Moderate	Rudner (1977c) Merz and Grossen (1979)
Difficult	Area Between ICC Curves	Conservative, 1000 per Group	Good Satisfactory Good	Rudner (1977c) Merz and Grossen (1979) Rudner, Getson and Knight (197
Simple	Significance of Chi-Square	100 or More per Group	Not Examined	

Each of these techniques will be discussed briefly. Although there is a wealth of literature dealing with these techniques, most of the material which will be discussed below is taken from the work of Rudner (1977a-d), Merz and Rudner (1978), Merz and Grossen (1979), and Rudner, Getson, and Knight (1979, 1980). The purpose of this discussion is to paraphrase, categorize, analyze, and synthesize the material presented in the works cited above. In addition to those categories included in Table 1 the assumptions and limitations of each method will be presented.

Analysis Of Variance (ANOVA)

Essentially this approach applies a technique for examining simultaneously the differences between many means. It begins with item scores; that is, the analysis involves the mean number of individuals responding to an item or the difficulty level (percent passing the item). Because the same individual takes all of the items, a repeated measures analysis of variance must be applied. The independent variables are group membership and items; the dependent variable is the numbers of individuals responding correctly. Because the analysis deals with percentages (the mean number passing an item), certain transformations or adjustments of the data are suggested to overcome problems encountered statistically when percentages are analyzed. The focus of the analysis is on the interaction between groups and items. Originally, the technique was coupled with a visual display of item difficulties for each pair of groups for which the analysis was done. This enabled the examiner to indicate which of the items interacted with group membership.

Later, Plake and Hoover (1977) developed a method where statistical contrasts between items could be made. The interaction contrast for each item within each group takes the form

$$\Psi_{ij} = \theta_{ij} - \theta_{i.} - \theta_{.j} + \theta_{..}$$

where θ_{ij} is the arcsine transformed item difficulty for the i^{th} item in the j^{th} group.

The error variance is

$$\sigma_{\Psi_{ij}} = \frac{q-1}{2q\,\tilde{n}_j}$$

where q is the number of items and

$\tilde{n}_j$ is the harmonic mean of the number of persons in the j^{th} group.

Employing this value, a significance test can be used to identify individual items which appear to be biased.

What the technique does is to contrast the proportions of each identified group attaining an item correct. If there is a significant item by group interaction, the plots in contrast analyses identify the biased items.

The technique is quite cumbersome and requires a great deal of computation. It does not consider the level of ability of individuals, nor does it consider differences in level achievement of varying groups. The major reason for discussing the method is its historical significance. The technique was first employed by Cardall and Coffman (1964). It was applied by Cleary and Hilton (1968) to a very interesting set of aptitude test scores (SAT). Other methods appear to function more accurately than this one, and it is not recommended for use at this time.

Transformed Item Difficulties (TID)

This method also begins with item scores. However, total scores are used as well. The procedure is to transform percent of correct responses (item p-values) to z values for each group independently; there are two methods of doing this; one was advocated by Angoff (1972) and used by Angoff (1972), Angoff and Ford (1974), Angoff and Modu (1973), and Echternacht (1974). The second was proposed by Rudner (1977c) and has been used by Rudner (1977c), Rudner, Getson, and Knight (1979), and Merz and Grossen (1979).

The first approach begins with item difficulties or p-values which are transformed for each sample independently through an inverse normal function and then to Δ values. Here $\Delta_{ij} = 13 + 4\, z_{ij}$

where z_{ij} is the inverse normal deviate for item i for group j.

The Δ_{ij} pairs are plotted on a scatterplot, and the distance (d) is used to indicate the size of bias.

$$d_i = |a\Delta_{i1} - \Delta_{i2} + b| / \sqrt{a^2 - 1}$$

where a is the slope of the regression line

b is the intercept,

Δ_{i1} is the Δ for item i for group 1, and

Δ_{i2} is the Δ for item i for group 2.

The second technique involves converting each of the item p-values to a z score using the following formula:

$$Z_{ij} = \frac{P_{ij} - P_{\cdot j}}{S_{pj}}$$

where P_{ij} is the percent of pupils in group j passing item i;

$p_{\cdot j}$ is the average percent correct for the whole test for group j_i, and

S_{pj} is the standard deviation of the percent correct for the whole test for group j.

Here

$$d = \frac{Z_{i1} - Z_{i2}}{\sqrt{2}}$$

The magnitude of the difference under both methods has been fixed at .75 by Rudner, although others have suggested 1.5.

Because these formulae may seem intimadating, it may be simpler to treat the concepts which underlie the method. In essence, TID couples plotting techniques along with finding the differences between percent of individuals responding correctly to an item in each group. The transformations are necessary because percentages have peculiar characteristics which make their comparability, especially under extreme cases,very difficult. By transforming the items to z-values and Δ-values, the statistical property of percentages is overcome. The method adjusts for the ability level of the two groups by converting in terms of distances from the mean of each group. This, of course, assumes that the group mean for the entire test adequately accounts for differences in ability for the whole group. The method does not take into consideration individual ability levels although it uses the group mean as an estimate of ability for the group as a whole. The procedure has been evaluated by Rudner (1977c)

Groome, M L (1977), Groome, W R (1977); Merz and Grossen (1979); Rudner, Getson, and Knight (1979); and Rudner, Getson, and Knight (1980). The method is simple to compute and easy to interpret. Among the seven it will be one recommended for use in the assessment of competency tests for bias. A detailed presentation of the computation will be presented later.

Point Biserial Correlation (PBIS)

The point biserial correlation approach was introduced by Green (1971, 1973, 1976). It has been used by Green and Draper (1972), and Ozenne, Van Gelder, and Cohen (1974). The approach begins by correlating the response of each individual to total score. Thus magnitudes of correlations for each item are computed for two groups at a time. Bias is defined by a difference which exceeds .20. Rudner, Getson, and Knight (1979) considered this a modification of the transformed item difficulty approach. Green and Draper (1972) regarded this method as analogous to adjusting means in the analysis of covariance, where adjustments to ability level are made automatically. However, in this approach both item discrimination and item difficulty influence item location on the regression plot; whereas, with the transformed item difficulty, percent passing alone determine the location.

Some of the major difficulties with the approach center on the magnitude of the correlations from right/wrong responses. In addition, because the items are correlated with total score, skewness in the distribution or restriction in variance minimize the correlation coefficient, as well. As the approach has been used, no significance test has been

employed; the magnitudes of difference have merely been compared to an arbitrary value, .20. The approach is straightforward and computationally fairly simple. In comparing this approach to others, however, Merz and Grossen (1979) found it to be among the least accurate in identifying biased items and it is not recommended for use.

Chi-Square (CHI)

This particular technique was developed by Scheuneman (1975, 1976). It relates item performance to total score performance by dividing the total score into a specific number of categories and tabulating the number of individuals within each total score category who got the item correct. Usually, the categories are quarters or fifths; that is, Scheuneman (1975) divided the total score distribution on the basis of the distribution of the smallest group. This approach would have to be adjusted to avoid having no one in one of the scoring categories. Rudner, Getson, and Knight (1979) suggest that the accuracy of this technique declines as the number of total score units is increased. They suggest using five categories per group. Merz and Grossen (1979) found eight categories to be satisfactory.

The technique tests whether examinees of the same ability level have the same proportion of correct responses regardless of cultural affiliation. An item is unbiased, then, if for all individuals in the same total score interval the proportion of correct responses is the same for the groups under consideration. A modified Chi-square is used to estimate the probability that the item is unbiased. The expected values for each cell

are obtained by the formula below:

$$E_{ij} = \frac{O_{.j}}{N_{.j}} N_{ij}$$

where $O_{.j}$ is the number of examinees in total score interval j responding correctly,

$N_{.j}$ is the total number of examinees in total score interval j, and

N_{ij} is the total number of examinees in group i and score interval j.

From this point the usual Chi-square is computed utilizing the formula

$$\chi_i^2 = \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where i is the item,

O_{ij} is the observed number in group j obtaining item i correct,

E_{ij} is the expected number in group j to get item i correct, and

$\sum_{j=1}^{k}$ directs one to sum from cell 1 to cell k.

The degrees of freedom to judge the significance of the Chi-square value obtained in the analysis has been in contention; a suggestion for the degrees of freedom will be made when this method is presented in detail with an example later.

Scheuneman considers this approach a specific application of item characteristic curve theory. Merz and Grossen (1979) and Rudner, Getson,

and Knight (1979) found this technique to be a reasonable method for identifying biased items.

Factor Analysis (FACTOR)

In factor analysis correct and incorrect responses to each item are correlated. This set or matrix of correlations is reduced to a smaller set of dimensions or traits which the test is presumed to measure. The traits or factors are named by the pattern with which items correlate or load on that factor. Essentially the pattern of factor loadings for correlations for different groups of individuals can be compared. If the pattern of loadings is similar for the two groups, there is evidence that the test items are functioning similarly for all groups concerned. If not, there is evidence that performance is influenced by different traits.

Rudner in Merz and Rudner (1978) described the mathematical model upon which this analysis is based as

$$F = Y_z R^{-1} A + E$$

where Y_z is a vector of subject responses in z score form

A is a matrix of factor loadings

F is a subject vector of factor scores

E is a vector of residual or error terms

R is the variable by variable correlation matrix

Three different approaches have been suggested to identify biased items with FACTOR. Green and Draper (1972) and Green (1976) advocated using Tucker's (1958) inter-battery factor analysis approach. In this method item variance is partioned into three segments: (1) Factors

common to each group, (2) Factors specific to each group, and (3) Residual or error variance. Here bias is determined by the variance attributable to each group individually. If that variance is small, items within the test are considered to be unbiased; if the proportion is large, the items are considered to be biased. Merz (1973, 1976) suggested two approaches in using FACTOR. The first couples factor analysis with analysis of variance. After factor analysis of item responses of all groups pooled together and computation of factor scores for each item, analysis of variance is employed to identify significant differences in the mean values of factor scores for each group. A factor is biased if there is a significant effect for group membership. Items which have major loadings or correlations with this factor are considered biased. The second method employs a set of group membership variables which are entered into the correlation and factor analysis. Items which correlate or have high factor loadings with group membership variables are considered biased.

To date there has been no literature found comparing the Green and Draper method with other methods. The methods suggested by Merz have been compared by Rudner (1977c) and by Merz and Grossen (1979). The accuracy of results was far from encouraging. This method performed less well than TID, CHI, or ICC approaches. In addition, the method is very difficult to conceptualize and requires a computer for analysis. For these reasons it is recommended for use by those engaged in research who have had considerable background with factor analytic procedures. It is not recommended for use with competency tests at this time.

Item Characteristic Curve (ICC)

ICC approaches are a part of a general methodology called latent trait theory. The latent traits are presumed examinee characteristics which influence performance on the items within a test. For purposes of ICC approaches four different traits are assumed to account for performance and enter into the analysis. These are θ --the ability level of the examinee, a--the discrimination of the item, b--the difficulty of the item, and c--the ability to guess. These four tratis are related to performance on individual items through the probability of responding correctly to the item.

Latent trait theory was introduced to educational measurement by Lord and Novick (1978), although the original work is much earlier (Birnbaum, 1958; Rasch, 1960). In considering ICC approaches it is important to recognize that different theorists advocate slightly different approaches. There are three models which are presently being used. The one parameter model works only with ability (θ), difficulty (a), and probability of correct response. This particular approach has been called the Rasch model and has been popularized by Wright and his co-workers. The second model or two parameter model adds discrimination to the list of traits or pararmeters examined. Finally, the three parameter model examines difficulty (a), discrimination (b), and guessing (c). It is called the Birnbaum model or three parameter model. Hambleton and others (1978) call attention to three important concepts which must be considered in the application of ICC or latent

trait theory; they are dimensionality, local independence, and item characteristic curves. Dimensionality requires that a test measures only one ability or is unidimensional. Local independence requires that a person's answering one item correctly is not affected by answering any other item correctly. Item characteristic curves are the tool used to relate the probability of obtaining a correct response to an item. The material which follows is adapted from Rudner, Getson, and Knight (1979). Birnbaum (1968) proposed a general cumulative logistic model for estimating parameters:

$$P(u_g = 1 \mid \theta_i) = c_g + (1 - c_g)1 + \exp\{[-1.7a_g(\theta_i - b_g)]\}^{-1}$$

where θ_i is examinee i's ability,

$P(u_g = 1 \mid \theta_i)$ is the probability of a correct response,

a_g is the discrimination of item g,

b is difficulty of item g, and

c_g is an estimate of guessing for item g.

The relationship between $\theta_i (-4 < \theta_i < 4)$ and $P(u_g = 1 \mid \theta_i)$ is shown in Figure 1, where $c_g = 0.2$, the lower asymptote; $b_g = 0.0$, the inflection point of the curve; and $a_g = 1.5$.

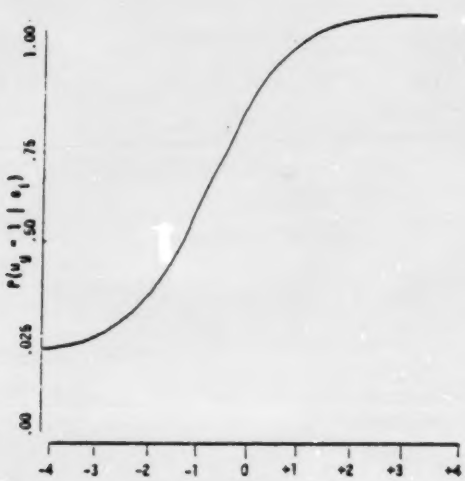

Figure 1. Hypothetical ICC

Figures 2, 3, and 4 illustrate the shapes of ICC curves influenced by differences in each of the three parameters.

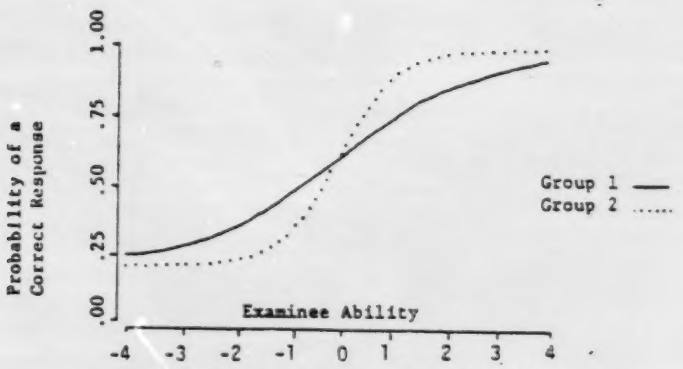

Figure 2. Bias Solely in Discrimination (a_g)

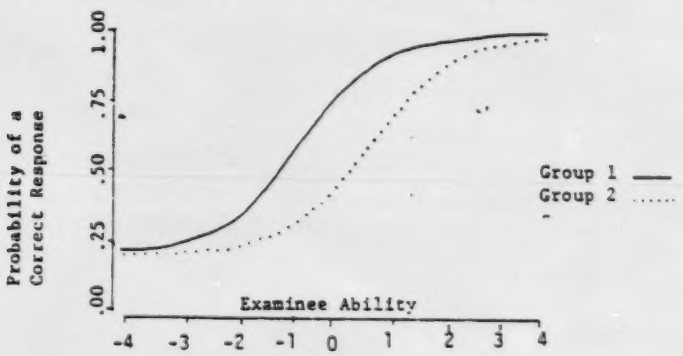

Figure 3. Bias Solely in Difficulty (b_g)

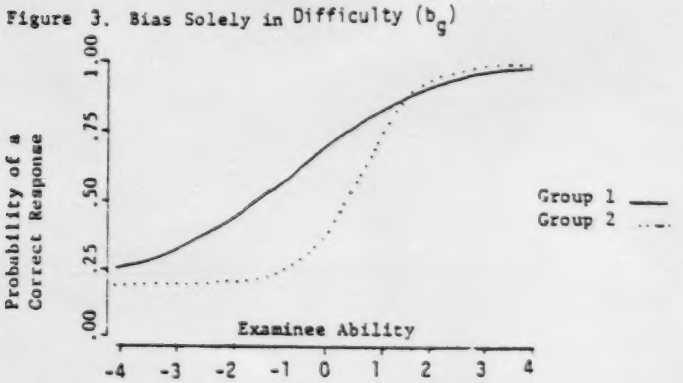

Figure 4. Bias in Both Difficulty (b_g) and Discrimination (a_g)

The focus of the analysis is on the difference in probability of responding correctly. The one parameter model examines only item difficulty with respect to ability and probability of correct response. The two parameter model examines item difficulty and item discrimination, and the three parameter model examines item difficulty, item discrimination, and guessing. The difference between ICC's for groups is the measure of bias.

Although the method is difficult to compute, computer programs are available to execute the complex mathematical operations in an efficient manner. The technique requires large numbers of examinees. One must have a good grasp of the mathematical model upon which the analysis is based. Nevertheless, the accuracy of the method recommends its use (Rudner, 1977c; Merz and Grossen, 1979; Rudner, Getson, and Knight, 1980). It is included among the methods recommended for use in detecting bias in test items.

Response Foil Analysis (RFA)

Response foil analysis or distractor analysis was introduced by Veale and Foreman (1975, 1976). With this method the distribution of responses to distractors for different groups are examined. If a group is disproportionately attracted to a distractor in comparison to other groups, this disproportionate attraction or pull might indicate bias. An example is indicated in Table 2. Here Group 1 has 63 percent passing the item (a p-value of .63), while Group 2 has 26 percent passing (a p-value of .26). The item has four options with response (b) being correct. The usual chi-square is computed for a three by two table. The value of chi-square is 12.58, significant beyond the .01 level. This shows that a disproportionate number of individuals in Group 2 are attracted to Response A . Thus, item writers and editors would need to investigate how Group 2 was attracted to this particular response.

Table 2

HYPOTHETICAL DISTRIBUTION OF DISTRACTOR CHOICES

	Response A		Response C		Response D		
	Observed	Expected	Observed	Expected	Observed	Expected	Total
Group 1	12	20.67	13	9.00	12	7.33	37
Group 2	50	41.33	14	18.00	10	14.67	74
Total	62		27		22		

$x^2 = 12.58$

$.001 < p < .01$

Summary

Of the seven methods presented, three are recommended for general use in detecting test item bias. Transformed item difficulty (TID) is probably the most simple of the methods to understand and to compute. Rudner, Getson, and Knight (1980) indicate that TID is sensitive to differences in item difficulty but not to differences in item discrimination. Chi-square (CHI) does not have the same problem with differences in item discrimination. It, too, is simple to compute, once individuals have been categorized into groups by total score. Having a computer available makes this process far easier. The three parameter item characteristic curve (ICC) method is quite complex to compute and to interpret. It requires larger numbers of examinees than either of the other two approaches. Yet, its accuracy and ability to examine three parameters recommend it where its complexity does not interfere with its use. Examples of the three methods are presented below.

Examples

Data Set Used

The examples included here are based on hypothetically generated data. Table 3 gives the total raw score distribution for two groups of 500 examinees on a 60 item test. The means of the two groups are within 2.14 raw score points of each other. Variances are close to each other, and reliabilities are both high. The differences between means, however, is statistically significant beyond the .01 level.

Table 3

TOTAL SCORE DISTRIBUTIONS FOR EXAMPLES

Score	Group 1 Frequency	Group 2 Frequency	Score	Group 1 Frequency	Group 2 Frequency
60	3	3	34	17	12
59	6	4	33	10	18
58	6	5	32	14	9
57	13	11	31	7	11
56	14	7	30	5	8
55	16	12	29	7	13
54	12	5	28	6	4
53	12	14	27	12	14
52	21	11	26	8	9
51	14	16	25	7	6
50	13	20	24	10	8
49	19	15	23	4	4
48	16	18	22	7	6
47	23	18	21	6	6
46	11	13	20	2	4
45	20	8	19	0	9
44	15	23	18	1	5
43	15	11	17	1	7
42	16	10	16	2	2
41	11	16	15	3	4
40	16	22	14	0	6
39	17	18	13	0	0
38	14	19	12	2	1
37	15	7	11	0	0
36	16	18	10	0	0
35	14	10	9	0	0

Table 3
(cont'd)

Score	Group 1 Frequency	Group 2 Frequency
8	0	0
7	0	0
6	1	0
5	0	0
4	0	0
3	0	0
2	0	0
1	0	0
0	0	0
Number	500	500
Mean	41.384	39.244
Standard Deviation	10.696	11.345
KR_{20}	0.917	0.923

$t = 3.07$

$.01 > p > .005$

The listing of item difficulties for each group is found in Table 4. The mean item difficulties are within 0.036 of each other. The difficulties for Group 1 range from 0.366 to 0.974; for Group 2, from 0.290 to 0.958.

Table 4

LISTING OF ITEM DIFFICULTIES
FOR EACH GROUP

Item	Group 1	Group 2	Item	Group 1	Group 2
1	0.940	0.948	31	0.366	0.512
2	0.790	0.826	32	0.890	0.958
3	0.776	0.840	33	0.680	0.454
4	0.396	0.570	34	0.474	0.568
5	0.678	0.634	35	0.974	0.906
6	0.872	0.690	36	0.614	0.556
7	0.584	0.334	37	0.836	0.906
8	0.624	0.432	38	0.758	0.602
9	0.850	0.874	39	0.356	0.382
10	0.430	0.522	40	0.846	0.792
11	0.720	0.786	41	0.958	0.826
12	0.792	0.562	42	0.544	0.590
13	0.664	0.660	43	0.554	0.378
14	0.738	0.594	44	0.770	0.708
15	0.596	0.590	45	0.600	0.516
16	0.884	0.792	46	0.896	0.864
17	0.762	0.930	47	0.718	0.648
18	0.768	0.706	48	0.646	0.566
19	0.844	0.866	49	0.898	0.930
20	0.398	0.466	50	0.586	0.290
21	0.618	0.744	51	0.364	0.480
22	0.760	0.708	52	0.760	0.710
23	0.794	0.632	53	0.620	0.586
24	0.832	0.846	54	0.616	0.646
25	0.866	0.876	55	0.542	0.644
26	0.560	0.412	56	0.602	0.556
27	0.476	0.406	57	0.658	0.594
28	0.796	0.818	58	0.796	0.664
29	0.526	0.466	59	0.454	0.572
30	0.886	0.672	60	0.788	0.638
			Mean	.690	.654
			Standard Deviation	.164	.170

Transformed Item Difficulty (TID)

Three basic steps are used. First, item difficulties are transformed to 2 scores; then, a distance measure is calculated. Finally, this distance measure is compared to a specific value. The example is calculated for item 7.

Steps	Example
1. Convert item p values to a z-score for each group separately. $z_{ij} = \frac{P_{ij} - P_{ij}}{S_{pj}}$	Group 1 $z_{71} = \frac{0.584 - 0.690}{0.164} = -0.646$ Group 2 $z_{72} = \frac{0.334 - 0.654}{0.170} = -1.882$
2. Convert z-scores to a distance function. $d = \frac{z_{i1} - z_{i2}}{\sqrt{2}}$	$d = \frac{(-0.646) - (-1.882)}{\sqrt{2}} = 0.874$
3. Compare the distance (d_i) to a given value (0.75) at Rudner's (1977c) suggestion.	$d_7 = 0.874$ $0.874 > 0.75$ Therefore, item 7 is biased.

Here item 7 is biased against Group 2 or in favor of Group 1. Now, careful inspection of item 7 will have to be made in order to determine what element or elements might bias the item. Very often bias in language accounts for differences. Since this is hypothetically generated data, no actual item exists, so no inspection can be made.

Chi-Square (CHI)

The distribution of total scores is broken down roughly into five catefories composed of 20 percent of the examinees by the distribution of the smaller group. In this case both groups are of equal size, so arbitrarily Group 1 was selected. Then, tabulations of the number right in each of the total score categories by group is made. Finally, chi-square is calculated.

Steps	Example				
	Score Range		N		O
1. Divide the total score into five categories, having roughly 20 percent of the smaller group in each category.	54 - 60	Group 1	47		32
		Group 2	70		61
		N_{ij}	117	O_{ij}	93
	50 - 53	Group 1	61		39
		Group 2	60		52
		N_{ij}	121	O_{ij}	91
	45 - 49	Group 1	72		31
		Group 2	67		67
		N_{ij}	139	O_{ij}	98
2. Tabulate the number Observed in each group.	38 - 44	Group 1	119		38
		Group 2	104		53
		N_{ij}	223	O_{ij}	91
	1 - 37	Group 1	201		27
		Group 2	177		59
		N_{ij}	370	O_{ij}	86

3. Compute the Expected number for each category using the formula

$$E_{ij} = \frac{O_{ij}}{N_{.j}} N_{ij}$$

where E_{ij} is the expected number in group i for interval j,

O_j is the observed number in interval j getting the item correct,

N_j is the total number in interval j, and

N_{ij} is the number in group i and score interval j.

Score Range		N	O	E	Cell Squar.
54 - 60	Group 1	47	32	37.36	0.769
	Group 2	70	61	55.64	0.51
	N_{ij}	117	O_{ij} 93		
50 - 53	Group 1	61	39	45.88	1.03
	Group 2	60	52	45.12	1.049
	N_{ij}	121	O_{ij} 91		
45 - 49	Group 1	72	31	43.83	3.756
	Group 2	67	67	54.17	3.03
	N_{ij}	139	O_{ij} 98		
38 - 44	Group 1	119	38	48.56	2.29
	Group 2	104	53	42.44	2.62
	N_{ij}	223	O_{ij} 91		
1 - 37	Group 1	201	27	45.73	7.67
	Group 2	177	59	40.27	8.711
	N_{ij}	378	O_{ij} 86		

$X^2 = 31.467$

$p < .001$

4. Compute Chi-Square, using the formula

$$x^2 = \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Here item 7 appears to overestimate the number in intervals 1-37, 38-44, and 45-59 for Group 1 and underestimate in the same intervals for Group 2. The Chi-square is significant.

Item Characteristic Curve (ICC)

ICC approaches relate the probability of responding correctly on an item, $P(u_g = 1|\theta_i)$, to ability on that item θ_i. Three parameters modify the shape of the ICC's; they are item discrimination (a_g), item difficulty (b_g), and guessing (c_g). The computer program used to obtain estimates of these parameters was prepared by Rudner (1977b). The measure used for the ability scale (θ_i) is defined by the ability variance in the sample of persons taking a given test. The program (PARAM) does this by computing the regressions of parameters of one group of examinees against the parameters of another group. ICC's will be identical when the test is unidimensional, contains independent items, and has error free parameter estimates.

The area between the equated ICC's is computed as the measure of bias. This area is computed by the formula

$$\theta_g = \int_{-\infty}^{\infty} [\,|P(u_g = 1|\theta_i) - P^1(u_g = 1|\theta_i)|\,]\, d\theta$$

where $P(u_g = 1|\theta_i)$ and $P^1(u_g = 1|\theta_i)$ are the equated ICC's for the groups.

PARAM approximates this value with the formula

$$\theta_g = \sum_{-5.0}^{+5.0} [\,|P(u_g = 1|\theta_i) - P^1(u_g = 1|\theta_i)\,]\, \Delta\theta$$

where $\Delta\theta = 0.005$

For item 7 the area between curves is 1.528, indicating that the item is biased.

Each of the three methods have computer programs available to do the calculations. Those used by the present investigator were written or adapted by Rudner (1976b). The only method which requires the use of a computer is the ICC approach. For both of the other approaches, hand calculation is possible, although arduous.

FAIR USE OF TESTS

Test fairness was defined as the equitable use of a test score in rendering judgment about an individual. The approaches involved are competing. By and large establishing that a test was used fairly has involved determining whether similar predictions from the test to an outside criterion could be obtained for each group to which the test was administered (Merz, 1978). When an outside criterion is not used fair use of tests is determined by comparing proportions passing from various groups of interest. Three major difficulties stand in the way of assessing the fair use of tests. First, the statistical methods available to perform analyses present conflicting approaches and assumptions (see JEM, 1976, 13:1). Second, identification and measurement of adequate criteria of success present logical problems which interact with the statistical methods employed to investigate fairness. Third, obtaining criterion data is extremely difficult and expensive.

Regression Approaches

While a great deal has been written on the various methods used to establish fair test use, there is a definite lack of clarity on which method to use because as Darlington (1976) points out many of the questions involved in the fair use of tests must be handled rationally before statistical methods can be selected and applied to the data. Merz (1978) summarized Petersen and Novick's (1976) critique of eight methods which use the prediction-regression approach.

The first method is called a regression approach and was used by Cleary (1968). With this approach a test is fair if applicants with the

highest predicted criterion scores are selected using separate regression equations for each group. This relationship is described by the equation

$$Y^* = a_1 + b_1 x_1^* = \ldots = a_g + b_g X_g^*$$

where Y^* is the minimum level of satisfactory critetion performance,

a_i is the intercept for group i,

b_i is the slope, and

X_i^* is the predictor pass score for group i.

A correction for different mean values is accomplished by using separate prediction equations. The method focuses on fairness to the individual rather than on fairness to the group. It is the most widely used approach to fair selection.

A second approach was introduced by Thorndike (1971 and has been called the constant ratio method. It defines as fair a test which selects individuals in such a way that the ratio of the proportion selected to the proportion successful after selection is the same for all subgroups. This method focuses on fairness to the group rather than fairness to the individual. It assumes that a constant ratio of success is reasonable for all subgroups. It is used where equity between groups is the central consideration and in situations where the differences between the means for predictors is different from the differences between means for the criterion.

The third method, called the equal risk method, was introduced by Einhorn and Bass (1971). It defines a test as fair if the test identifies those persons whose prediction of success can be made with a specified degree of confidence. The cutting scores or passing scores are selected in such a way that the probability of success is equal for each group. This

is done by adjusting the criterion passing score with a confidence band using this formula:

$$Y = Y_c - z_p (S_{y.x})$$

where z_p is a z-score which can be designated by the desired degree of risk,

Y_c is the criterion passing score, and

$S_{x.y}$ is the standard error of estimate.

The approach allows separate cutoff or passing scores for each group if the standard errors for the groups are different. It focuses on group equity rather than individual fairness.

The fourth method has been called the culture-modified criterion method (Darlington, 1971). It defines as fair a test which selects individuals with the highest predicted scores, using subjectively modified criterion variables for each group. The modification of criterion variables is done to make the test culturally optimal; that is, (1) a subjective policy level question is answered concerning the optimum balance between performance and cultural factors, and (2) an empirical relationship between a test and a culture-modified variable (Y-kC) is established. The regression method equation is modified in this fashion

$$(Y-k_iC_i) = a_i + b_iX_i^*$$

where Y is the criterion,

k_i is constant, subjective value judgment on the part of the decision maker,

C_i is an applicant's group membership,

a_i is the intercept for group i,

b_i is the slope for group i, and

X_i^* is the predictor pass score for group i.

The fifth approach is called the conditional probability method (Cole, 1973). It defines a test as fair if all examinees who are capable of being successful are guaranteed an equal opportunity of being selected regardless of the group to which they belong. The method looks for equity to the group and like the constant ratio method yields a constant to which probabilities across groups must be equal.

The sixth method is the equal probability method introduced by Linn (1973). It defines a test as fair if all examinees who are selected are guaranteed an equal chance of being successful regardless of group membership. This method like the conditional probability method and constant ratio method corrects on the marginal distribution of the predictor; that is, it yields a constant for all groups to which the probability must be equal. For each of these three methods to work it must be reasonable to expect all groups to perform equally well on the predictor.

The seventh method is the threshold utility approach introduced by Gross and Su (1975). It states that a test is fair if individuals are selected by predictor passing scores which maximize the expected utility of the selection process. This is done by assigning different utilties to different groups. The term utility can be roughly equated with success. The method focuses on successful prediction by seeking public statement of utilities and, then, maximizing the likelihood of the utility for a given group.

A complete discussion of these methods may be found in the JEM, 1976, 13: 1, in Merz (1978), and in Merz and Rudner (1978). The important issue in all of these methods is that they demand an outside criterion of success.

Thus, if the methods are to be used, some external criterion of success must be designated and data about performance on that criterion must be collected. This presents a large problem for competency tests because many of the desired outcomes which indicate the success of compulsory education are evident only after schooling is completed. These outcomes are, in fact, the criteria of success on which data must be collected. In addition, competency testing was introduced to verify that diplomas were granted to those individuals who possess sufficient academic achievement to merit a diploma. To verify the competency test another criterion independent of the test and the diploma would be needed. Designating such a criterion would be difficult, indeed. A very likely criterion which might be under the control of the testing authority would be another test. Again, finding such a test and administering it presents a large number of operational problems.

Adverse Impact Approach

One possible answer to this dilema is to employ a method which does not require an outside criterion. To do this one might borrow from the Equal Employment Opportunity Guidelines (1978) a procedure called the 80 percent rule. Here the concept of adverse impact is introduced. In essence, adverse impact on a group is viewed as a passing rate less than 80 percent of the group with the highest passing rate. The example below will demonstrate how this works.

Table 5

HYPOTHETICAL APPLICATION OF THE 80 PERCENT RULE NO ADVERSE IMPACT

	Group 1	Group 2	Group 3
Number in Group	100	50	30
Number Passing	75	30	20
Passing Rate	75%	60%	66.6%
Passing Rate by 80 Percent Rule	75%	60%	60%
Number Needed by 80 Percent Rule	75	30	18

There are three groups having 100, 50, and 30 examinees respectively in each group. The highest passing rate is for Group 1 which has 75 of the 100 individuals passing for a rate of 75 percent. Using the 80 percent rule Group 2 would need 80 percent of 75 percent, or 60 percent. This equals 30 people. By the same token Group 3 would need 60 percent or 18 people if there were no adverse impact. In the example there is no adverse impact because Groups 2 and 3 have passing rates which equal or exceed the rate established by the 80 percent rule.

A second example will show adverse impact.

Table 6

HYPOTHETICAL APPLICATION
OF THE 80 PERCENT RULE
ADVERSE IMPACT

	Group 1	Group 2	Group 3
Number in Group	100	50	30
Number Passing	80	30	20
Passing Rate	80%	60%	66.6%
Passing Rate by 80 Percent Rule	80%	64%	64%
Number Needed by 80 Percent Rule	80	32	19

Here the group size has remained the same but the number passing in Group 1 has been raised to 80, yielding a passing rate of 64 percent for Groups 2 and 3 by the 80 percent rule. However, Group 2 has a lower passing rate, indicating adverse impact. Group 3 has a higher passing rate, indicating no adverse impact.

While this approach seems simplistic, it provides a method which has been accepted by the courts in employment cases. One caution, however, is

that this is only an indication that a problem may exist. It is possible in the second example that Group 2 does, in fact, have a poorer level of achievement than either Group 1 or Group 3. In this case, the test could be fair and simply indicate that Group 2 needs additional resources to meet the requirements for competence. This statement does not beg the question; it merely points out that more than one possibility exists when groups are observed to perform differently. One must determine whether a group has been fairly treated.

One indication which may be used to answer questions about the fairness of a measure is the performance history of a group. One may ask whether members of a group consistently, over a long period of time, have a lower passing rate than members of other groups. If the history of lower passing rates is present many questions must be asked about the causes for this situation and what has been done to remedy the situation before one asks whether or not the test has been used fairly. In examining the test itself one must inspect directions for clarity, accuracy, and ease of understanding. The method of marking must be carefully inspected to rule out mechanical causes for observed differences. Finally, test items themselves must be inspected for bias. This last possibility may make the whole process seem tautological; however, it is entirely possible that the presence of biased items may, in fact, account for the differences among groups in mean performance on a test.

RECOMMENDATIONS

In the case of competency testing it is clear that both bias and fair use of tests must be examined. Because bias is a concern primarily during the test construction phase, all items in the item pool must have some estimate of the effect of group membership on item performance. Fair use of a test must be reevaluated as conditions change; that is, as the pupil population changes, as the test is edited or reformatted, as new criteria for successful performance are added. In short, the task of assuring that test items are unbiased estimates of performance and that the test is being fairly used are ongoing tasks which must be part and parcel of any testing program.

Bias In Test Items

During the test construction phase it is recommended that one of three methods be employed:

(1) Transformed Item Difficulty,

(2) Chi-square, and

(3) Three Parameter Item Characteristic Curve.

Transformed item difficulty is appropriate when differences in difficulty not differences in discrimination are the question. The method can be computed simply with the aid of a desk calculator after item difficulties have been calculated. The chi-square technique can be used under almost all conditions. It can be computed simply with the aid of a desk calculator after the numbers of individuals responding correctly in specific total score ranges have been tabulated. To use this method it is necessary to arrange items into tests. Hence, it is appropriate to use the method during the trial phase of intact tests. The three parameter item

characteristic curve approach can be used under the same conditions that the chi-square approach can be used. Its chief limitations are that computer facilities will be necessary to perform the analysis and special training with latent trait models is a prerequisite for understanding the method and the analysis.

One should be aware that in all three cases there will probably be an overidentification of the number of biased items and that a small number of biased items will not be identified (Merz and Grossen 1978). This is the case with all statistical methods and models. There is a margin of error, and that margin of error must be taken into account.

The way this may be done is to examine logically the results of the analysis. Then, the test items themselves, must be examined to account for the findings. Finally, the test, as a whole, must be critically evaluated to examine for factors such as those listed in Wargo and Green (1978) and the material developed by Educational Testing Service for the California State Department of Education Proficiency Testing Program (1979). Statistical analysis is never a substitute for logic. For statistical analysis to function properly, the logic behind test construction and test utilization must be quite good. Thus statistical methodology must be coupled with good planning and careful item construction.

Fair Use Of Tests

In examining tests for fair use the first decision which must be made is whether or not an external criterion measure will be employed. If an external criterion is available, it must be identified and measures for success on it must be obtained. In addition, an acceptable level of performance on the competency measure must be established. Then, the

competency test itself must be administered, and an acceptable level of performance or a passing score must be established. The method used to analyze the data must be selected. First, all of the methods suggested are prediction-regression methods. However, each compensates for the adverse affect of specific group membership in a different way. Therefore, the type and amount of compensation made for group membership must be clearly stated at the outset so that appropriate adjustments may be made either to the competency test or the criterion measure.

If no outside criterion is to be used, the 80 percent rule can be applied to examine for adverse impact. If either adverse impact by the 80 percent rule or some indication that the test unfairly measures success on the criterion employing a regression model, the items in the test itself must be examined by one or more of the methods indicated in the section on bias to determine if test item bias contributes to the adverse impact or lack of fairness. If it does not, other areas must be examined to assure that factors extraneous to measurement are not contributing to differential performance. Other areas which must be examined are the directions which must be carefully written to indicate clearly to the examinee what he/she is to do. Systems for responding to items must minimize error of responding; that is, they must be appropriate to age level and test taking ability of the individuals for whom the test is designed. Scoring systems must be objective and free from systematic error. All three of these areas must be carefully considered if a test is judged to be unfair.

The whole issue of establishing a passing score must be considered, as well. While it is a confounding element in proficiency testing, it must be considered as a topic separate from bias and fairness. Yet, the topic of passing scores must be mentioned as an area which must be considered carefully.

Conclusion

The issue of assessing bias in fairness in tests is quite complex. No test, measurement, or observation will ever be free from error. What is required, however, is that a reasonable effort be made to avoid and account for systematic error introduced into the assessment process. This may be done by adopting and following a set of reasonable procedures to examine test item pools for bias and intact tests for fairness. Employing measurement techniques to assess for competency does not free one from the responsibility of making a decision. Too often, a score is used as the final decision. Scores and other information are simply evidence which must be used in the process of making a judgment. To use data such as test scores in making a judgment one must assure the individual being judged that the scores are as unbiased an estimate of the person's true performance as possible and that the decision process employing test data will give the person a fair chance to demonstrate competence.

References

Angoff, W. H. A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, May 1972.

Angoff, W. H., & Ford, S.F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-105.

Angoff, W. H., & Modu, C. C. Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. New York: College Entrance Examination Board, 1973.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novik, Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968, Chapters 17-20.

Cardall, C., & Coffman, W. R. A method for comparing performance of different groups on the items in a test. (RM 64-61) Princeton: Educational Testing Service, 1964.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cleary, T. A., & Hilton, T. L. An investigation into item bias. Educational and Psychological Measurement, 1968, 8, 61-75.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Darlington, R. B. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Darlington, R. B. A defencse of "rational" personnel selection, and two new methods. Journal of Educational Measurement, 1976, 1, 43-52.

Echternacht, G. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Educational Testing Service. Handbook for Proficiency Assessment. Prepared for California State Department of Education, 1979.

Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 1971, 75, 261-269.

Equal Employment Opportunity Commission. Uniform Guidelines on Employee Selection Practices, 1978.

Green, D. R. Racial and ethnic bias in test construction. USOE Contract No. OEC-9-70-58 (0057), CTB/McGraw-Hill, Monerey, 1971.

Green, D. R. Racial and ethnic bias in achievement tests and what to do about it. Monterey, CA: CTB/McGraw-Hill, 1973

Green, D. R. Reducing bias in achievement tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Green, D. R., & Draper, J. F. Exploratory Studies of Bias in Achievement Tests. Monterey: CTB/McGraw-Hill, 1972.

Groome, M. L. A comparison of two approaches for item bias identification. Unpublished masters thesis. Sacramento: California State University, Sacramento, 1977.

Groome, W. R. Item bias: a comparison of two approaches. Unpublished maters thesis. Sacramento: California State University, Sacramento, 1977.

Gross, A. L., & Su, W. Defining a "fair" or "unbiased" selection model: A question of utilities. Journal of Applied Psychology, 1975, 60, 345-351.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Ironson, G. A comparative analysis of several methods of assessing item bias. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Ironson, G. A comparative analysis of several methods of assessing item bias. Journal of Educational Measurement, 1979, 16(4).

Journal of Educational Measurement, 1976, 13:1.

Journal of Educational Measurement, 1977, 14:2

Larry P. v. Wilson Riles et al, Opinion, U. S. District Court for the Northern District of California, 1979.

Linn, R. L. Fair Test use in slection. Review of Educational Research, 1973, 43, 139-161.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores (2nd ed.) Reading MA: Addison-Wesley, 1974.

Merz, W. R. Factor analysis as a technique in analyzing test item bias. Paper presented at the annual meeting of the California Educational Research Association, Los Angeles, CA, 1973.

Merz, W. R. Estimating bias in test items utilizing principal components analysis and the general linear solution. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Merz, W. R. Test fairness and test bias: A review of procedures. Paper presented at the USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Program Evaluation, Reston, VA, May 1976.

Merz, W. R. Test fairness and test bias: a review of procedures. In M. J. Wargo & D. R. Green, Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey: CTB/McGraw-Hill, 1978.

Merz, W. R., & Grossen, N. E. An empirical investigation of six methods for examining test item bias. Unpublished final report, Grant NIE-6-78-0067, Calfornia State University, Sacramento, 1978.

Merz, W. R. & Rudner, L. M. Bias in testing: a presentation of selected methods. A paper as part of a co-sponsored symposium, Empirical Evaluations of Various Approaches for Identifying Biased Test Item(s), at the annual meeting of the American Educational Research Association and the National Council for Measurement in Education, Toronto, March 1978.

Ozenne, D. G., Van Gelder, N. C., & Cohen, A. J. Energizing School Aid Act (ESAT) National Evaluation, Achievement Test Restandardization. Santa Monica, CA: Systems Development Corporation, 1974.

Petersen. N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Plake, B. S., & Hoover, H. D. An analytical method of identifying biased test items. Paper presented at the annual meeting of the American Educational Research Association. New York. April 1977.

Rasch, G. Probablistic models for intelligence and attainment tests. Copenhagen: Nilsen and Lydiche (for Denmark's Paedagogich Institute), 1960.

Review of Educational Research, 1979, 48:4.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977a.

Rudner, L. M. A closer look at latent trait parameter invariance. Paper presented at the annual meeting of the New England Educational Research Organization, Manchester, NH, May 1977b.

Rudner, L. M. An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, The Catholic University of America, 1977c.

Rudner, L. M. Using standard tests with the hearing impaired: The problem of item bias. Volta Review, 1978, 80, 31-40.

Rudner, L. M., Getson, P. R., & Knight, D. L. The effect of various test and item properties on five approaches to biased item detection. Paper presented at the annual meeting of the National Council on Educational Measurement, San Francisco, April 1979b.

Rudner, L. M., Getson, P. R., and Knight, D. L. A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 1980, (in press).

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., April 1975.

Scheuneman, J. A procedure for evaluating item bias in the absence of an outside criterion. Paper presented at the annual meeting of the American Educational Research Association, April 1976.

Stanley, J. C. & Hopkins, K. D. Educational and psychological measurement and evaluation. 4th ed. Englewood Cliffs, N. J.: Prentice-Hill, 1972.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Tucker, L. R. An interbattery method of factor analysis. Psychometrika, 1958, 23, 111-136.

Veale, J. R., & Foreman, D. I. Cultural validity of items and tests: A new approach. Score Technical Report, Iowa City, Iowa: Westinghouse Learning Corporation/Measurement Research Center, 1975.

Veale, J. R. & Foreman, D. I. Cultural variation in criterion-referenced test items: a global item analysis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, 1976.

Wargo, M. J., & Green, D. R. (Eds) Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey: CTB/McGra-Hill, 1978.

CALIFORNIA STATE UNIVERSITY, SACRAMENTO
6000 J STREET, SACRAMENTO, CALIFORNIA 95819

SCHOOL OF EDUCATION
DEPARTMENT OF BEHAVIORAL SCIENCES

April 10, 1980

Dr. Daniel Antonopolos
Basic Skills Division
National Institute of Education
18th & M Streets, N.W.
Washington, D.C. 20208

Subject: Grant: NIE-G-78-0067
Project: 8-0425

Dear Dr. Antonopolos:

I wish to add an errata sheet to the final report of our item bias grant. Through an oversight material provided by Dr. Lawrence M. Rudner was included without citation in our final report. I wish to include the errata sheet with the final report. Fifteen copies of the errata sheet are enclosed for inclusion with the final report.

I regret any inconvenience that this might cause; however, in fairness to Dr. Rudner his work should be appropriately cited.

Sincerely,

William R. Merz, Ph.D.
Associate Professor
Department of Behavioral Sciences

WRM:jg
cc: L. Rudner
Enclosures: 15

Errata

Material included on pages 24, 26, 28, 29, 30, and 31 were taken from work by Rudner in Merz and Rudner (1978) and Rudner, Getson, and Knight (1979).

Merz, W. R. & Rudner, L. M. Bias in testing: a presentation of selective methods. A paper as part of a co-sponsored symposium, Empirical Evaluations of Various Approaches for Identifying Biased Test Item(s), at the annual meeting of the American Educational Research Association and the National Council for Measurement in Education, Toronto, March 1978.

Rudner, L. M., Getson, P. R., & Knight, D. L. The effect of various test and item properties on five approaches to biased item detection. Paper presented at the annual meeting of the National Council on Educational Measurement, San Francisco, April, 1979.